

This listing of claims will replace all prior versions, and listings, of claims in the application:

Claims 1-16 (canceled)

1   Claim 17 (currently amended): A system for building a  
2   lexicon for use in capitalization correction for  
3   unstructured excerpts, comprising:  
4         a ripper adapted to assemble a list of word sets  
5         from unstructured content, at least one of the word  
6         sets comprising a word and at least two non-standard  
7         capitalization variations for the word; and  
8         an aggregator adapted to aggregate at least one  
9         of the each word sets set, the aggregator including  
10            an analyzer adapted to identify non-standard  
11            capitalization variations based on at least one  
12            criteria; and  
13            a non-standard capitalization selector  
14            adapted to select at least one of the identified  
15            non-standard capitalization variations within one  
16            of the at least one word sets identified word  
17            set, and adding the selected at least one of the  
18            identified non-standard capitalization variations  
19            to the lexicon, wherein the lexicon includes  
20            records, each record including a word, wherein  
21            the lexicon is indexed by the words included in  
22            the records, and wherein at least one of the  
23            records includes more than one non-standard  
24            capitalization variation.

1 Claim 18 (previously presented): A system according  
2 to Claim 17, further comprising:

3 a tokenizer adapted to tokenize the excerpt into  
4 the one or more words and one or more punctuation  
5 marks.

1 Claim 19 (original): A system according to Claim 18,  
2 wherein hyphenated words are split into a plurality of  
3 the words.

Claim 20 (canceled)

1 Claim 21 (previously presented): A system according  
2 to Claim 17, wherein at least one of the non-standard  
3 capitalization variations occurs in an excerpt having  
4 fewer than half of individual letters provided in  
5 uppercase.

1 Claim 22 (previously presented): A system according  
2 to Claim 17, further comprising:  
3 a normalizer adapted to normalize a plurality of  
4 the words extracted relative to a source of the  
5 unstructured excerpt.

1 Claim 23 (previously presented): A system according  
2 to Claim 17, wherein non-standard capitalization  
3 variations that are identified based on one or more  
4 criteria comprise only those non-standard  
5 capitalization variations having at least four  
6 occurrences.

1 Claim 24 (previously presented): A system according  
2 to Claim 17, wherein at least one of the non-standard  
3 capitalization variations has any individual letter  
4 other than the first individual letter provided in  
5 uppercase.

Claim 25 (canceled)

1 Claim 26 (previously presented): A system according  
2 to Claim 17, further comprising:  
3 a validator adapted to apply implicit rules for  
4 capitalization, and skipping each of the non-standard  
5 capitalization variations subject to at least one such  
6 implicit rule.

1 Claim 27 (previously presented): A system according  
2 to Claim 26, wherein the implicit rules comprise  
3 skipping each of the non-standard capitalization  
4 variations based on position within a sentence or  
5 phrase.

1 Claim 28 (previously presented): A system according  
2 to Claim 26, wherein the implicit rules comprise at  
3 least one of (A) the non-standard capitalization  
4 variation being a number, (B) the non-standard  
5 capitalization variation having no vowels, and (C) the  
6 non-standard capitalization variation constituting at  
7 least one of an article, conjunction and preposition.

1 Claim 29 (previously presented): A system according  
2 to Claim 26, wherein the implicit rules comprise

3 normalizing a number of occurrences for each of the  
4 non-standard capitalization variations relative to a  
5 source of the non-standard capitalization variations.

1 Claim 30 (previously presented): A system according  
2 to Claim 26, wherein each of the word sets includes a  
3 word and at least one non-standard capitalization  
4 variation, each of the at least one non-standard  
5 capitalization variation including a frequency of  
6 occurrence count.

1 Claim 31 (original): A system according to Claim 17,  
2 further comprising:  
3 a hash table maintaining the lexicon.

1 Claim 32 (previously presented): A system according  
2 to Claim 31,  
3 wherein the hash table is indexed by words.

Claims 33-50 (canceled)

1 Claim 51 (currently amended): A computer-implemented  
2 method comprising:  
3 a) generating a plurality of word sets from a text  
4 corpus, each at least one of the words sets  
5 including  
6 - a word identified from the text corpus,  
7 - at least one non-standard capitalization  
8 variation of the word included in the word set,  
9 and

10           - a frequency of occurrence of each of the at  
11           least one non-standard capitalization variation  
12           of the word included in the word set; and  
13           b) generating a lexicon using the generated  
14           plurality of word sets, wherein the lexicon  
15           includes, for each of a plurality of words, at least  
16           one capitalization variation identified using at  
17           least one criteria, wherein at least one of the  
18           words of the lexicon includes more than one  
19           non-standard capitalization variation identified  
20           using the at least one criteria; and  
21           c) storing the generated lexicon.

1       Claim 52 (currently amended): The computer-implemented  
2       method of claim 51 wherein a non-standard capitalization  
3       variation is identified using the at least one criteria  
4       only if it occurs at least four times in the text corpus.

1       Claim 53 (currently amended): The computer-implemented  
2       method of claim 51 further comprising:  
3           de) accepting a word having a capitalization  
4           defining which, if any, of the characters of the  
5           word are capitalized; and  
6           ed) performing a capitalization correction function  
7           on the word using the generated lexicon.

1       Claim 54 (currently amended): The computer-implemented  
2       method of claim 53 wherein the act of performing a  
3       capitalization correction function includes  
4           - determining if the capitalization of the  
5           word matches a capitalization variation in the  
6           lexicon, and

7           - not changing the capitalization of the word  
8        if it was determined to match a capitalization  
9        variation in the lexicon.

1   Claim 55 (currently amended): The computer-implemented  
2   method of claim 53 wherein the act of performing a  
3   capitalization correction function includes  
4           - determining if the capitalization of the  
5        word matches a non-standard capitalization  
6        variation in the lexicon, which non-standard  
7        capitalization variation meets a frequency  
8        criteria, and  
9           - not changing the capitalization of the word  
10      if it was determined to match a non-standard  
11      capitalization variation in the lexicon.

1   Claim 56 (currently amended): Apparatus comprising:  
2       a) means for generating a plurality of word sets  
3       from a text corpus, at least one each of the word  
4       words sets including  
5           - a word identified from the text corpus,  
6           - at least one non-standard capitalization  
7        variation of the word included in the word set,  
8       and  
9           - a frequency of occurrence of each of the at  
10      least one non-standard capitalization variation  
11      of the word included in the word set; and  
12       b) means for generating a lexicon using the  
13       generated plurality of word sets, wherein the  
14       lexicon includes, for each of a plurality of words,  
15       at least one capitalization variation identified  
16       using at least one criteria, wherein at least one of

17       the words of the lexicon includes more than one  
18       non-standard capitalization variation identified  
19       using the at least one criteria.

1       Claim 57 (currently amended): The apparatus of claim 56  
2       wherein a non-standard capitalization variation is  
3       identified using the at least one criteria only if it  
4       occurs at least four times in the text corpus.

1       Claim 58 (previously presented): The apparatus of claim  
2       56 further comprising:  
3           c) means for accepting a word having a  
4           capitalization defining which, if any, of the  
5           characters of the word are capitalized; and  
6           d) means for performing a capitalization correction  
7           function on the word using the generated lexicon.

1       Claim 59 (previously presented): The apparatus of claim  
2       58 wherein the means for performing a capitalization  
3       correction function  
4           - determine if the capitalization of the word  
5           matches a capitalization variation in the  
6           lexicon, and  
7           - do not change the capitalization of the word  
8           if it was determined to match a capitalization  
9           variation in the lexicon.

1       Claim 60 (previously presented): The apparatus of claim  
2       58 wherein the means for performing a capitalization  
3       correction function  
4           - determine if the capitalization of the word  
5           matches a capitalization variation in the

6           lexicon, which capitalization variation meets a  
7           frequency criteria, and  
8           - do not change the capitalization of the word  
9           if it was determined to match a capitalization  
10          variation in the lexicon.